Supplemental - Harnessing the Conditioning Sensorium for Improved Image Translation

Cooper Nederhood[1,3], Nicholas Kolkin[2], Deqing Fu[1,3], and Jason Salavon[1,3]

[1]The University of Chicago [2]Toyota Technological Institute at Chicago [3]Jason Salavon Studio

{cnederhood, deqing, jsalavon}@uchicago.edu, {nick.kolkin}@ttic.edu

1. Additional Implementation Details

Content Encoder: In the original SPADE architecture, the raw content representation is projected to an embedding space and then convolved to produce the modulation parameters [2]. Sensorium introduces a Content Encoder which builds a pyramid of features maps that replace the downsampled raw content representation. Each feature map emitted from the Content Encoder has channel dimension of 128, which is chosen to match the channel dimension of the embedding space in [2]. The Content Encoder does not emit feature maps for resolutions 64, 128, and 256 as the feature maps at these resolutions will come from Style/Content Fusion modules. See Figure 1 for further details.

Style Encoder: The Style Encoder is domain specific and largely follows the encoder design from [1]. See Figure 2 for further details.

Style/Content Fusion: The original SPADE architecture only incorporates style information once at the base of the generator. To allow for greater style control we localize the global style information by fusing the style representation with the highest resolution content representation, $\bar{c_a}^{(32)}$ [2]. See Figure 3 for further details.

Generator: We have a single generator for all domains which learns to synthesize an image given a pyramid of spatial varying feature maps containing all style and content information. Each incoming feature map has channel dimension of 128, which matches the dimension of the embedding space in the original SPADE modules [2]. See Figure 4 for further details.

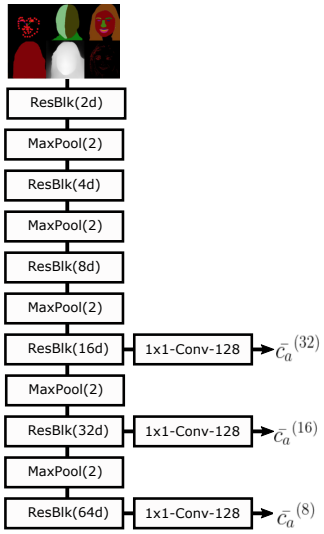

Figure 1: The Content Encoder model is shared across all domains and consists of a series of residual blocks and max pooling layers. The number of channels is a factor of d which varies based on the choice of conditioning. At each residual block, we cap the channel dimension at 512. Assuming a 256x256 input image, the feature maps of resolution 32, 16, and 8 are each passed through a convolutional layer and then emitted to form the hiearchical content representation $\bar{c_a} = \{\bar{c_a}^{(r)} \mid r \in \{8, 16, 32\}\}$. The term 1x1-Conv-128 denotes a 1-by-1 convolution with 128 convolutional filters.

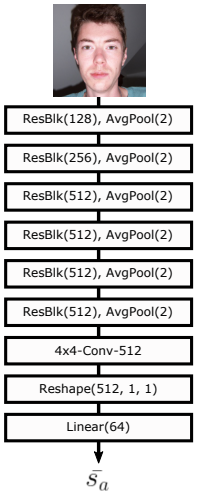

Figure 2: The Style Encoder model is domain-specific and applies a series of residual blocks and average pooling layers until the image resolution is reduced to 4x4 upon which another convolution fully reduces the spatial dimension and we apply one final linear layer.

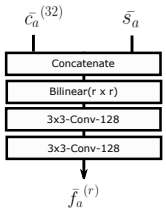

Figure 3: The Style/Content Fusion module is a simple way to combine the spatially varying content feature maps from $\bar{c_a}$ with the global representation of style from $\bar{s_a}$. Regardless of the current spatial dimension, r, the Style/Content Fusion model spatially replicates the global style $\bar{s_a}$ with $\bar{c_a}^{(32)}$, concatenates, upsamples to the desired resolution and then applies a series of convolutions.

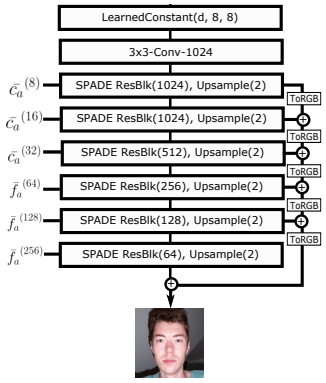

Figure 4: The Generator begins from a learned constant and applies a series of SPADE residual blocks. While in the original SPADE generator the scale and bias parameters are functions of the raw conditioning, in Sensorium the input to the SPADE layers is some feature map from the Content Encoder or a Content/Fusion module. At low resolutions the SPADE scale and bias parameters are functions of the content representation, $\bar{c_a}$, while at higher resolutions they are functions of a fused content/style representation. We add output skip connections, denoted by ToRGB, which is simply a 1-by-1 convolution with 3 feature maps and an upsampling to allow for element-wise addition with the next skip connection.

Figure 5: Additional examples of Sensorium's output for FFHQ-Wild translations. Sensorium performs three synthesis tasks: reconstruction (along the diagonal); within domain style transfer (in the main diagonal blocks); and domain translation (between gender in the off-diagonal blocks). Further, image quality is robust to changes in facial pose, geometric transformations of hair, and changes in camera zoom. Finally we note that Sensorium gracefully handles additional faces in the frame.

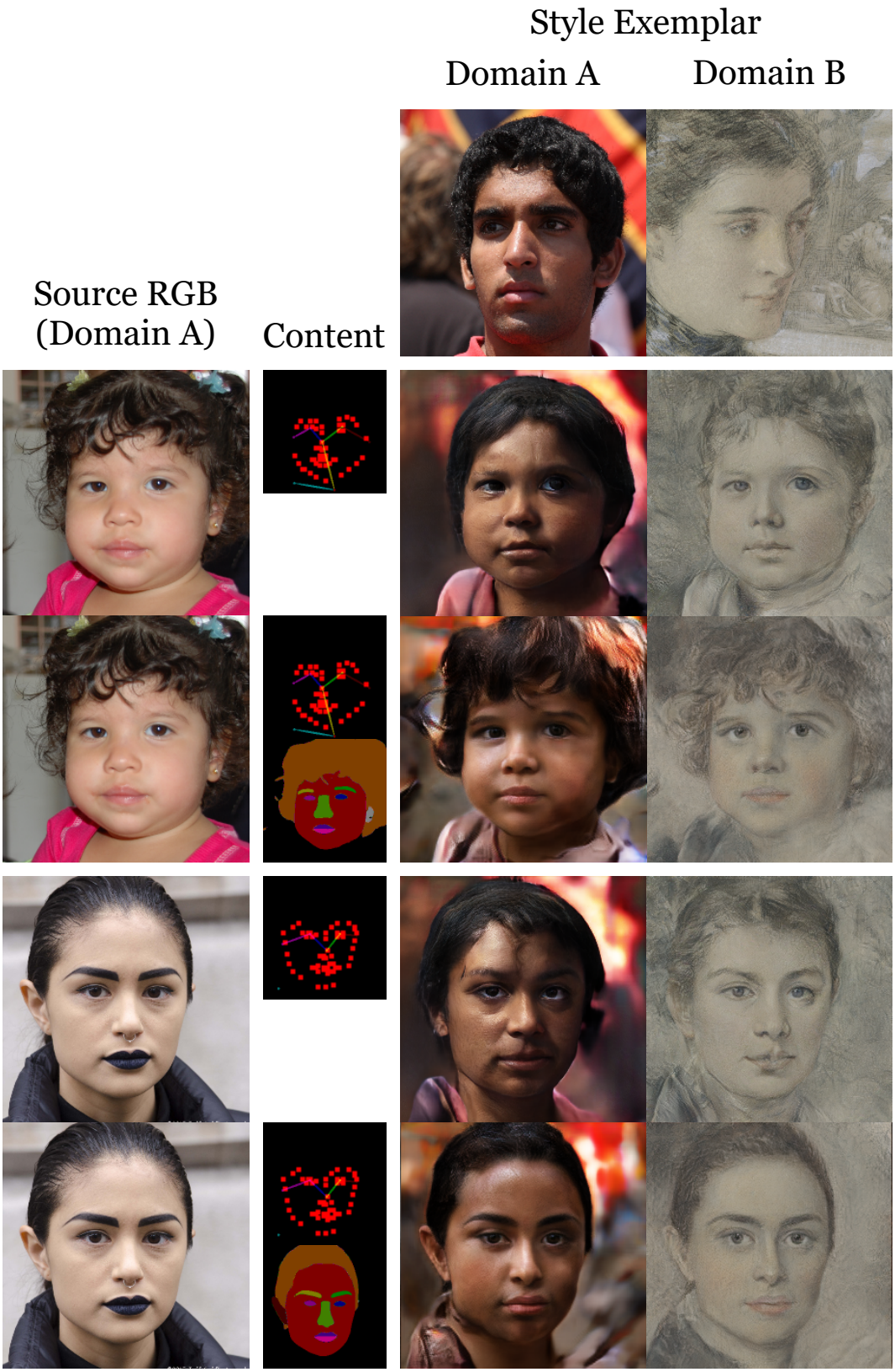

Figure 6: We show how different choices of conditioning allow for control in the synthesized image. When translating between human face styles the exact content representation is shaped by the user's preferences.

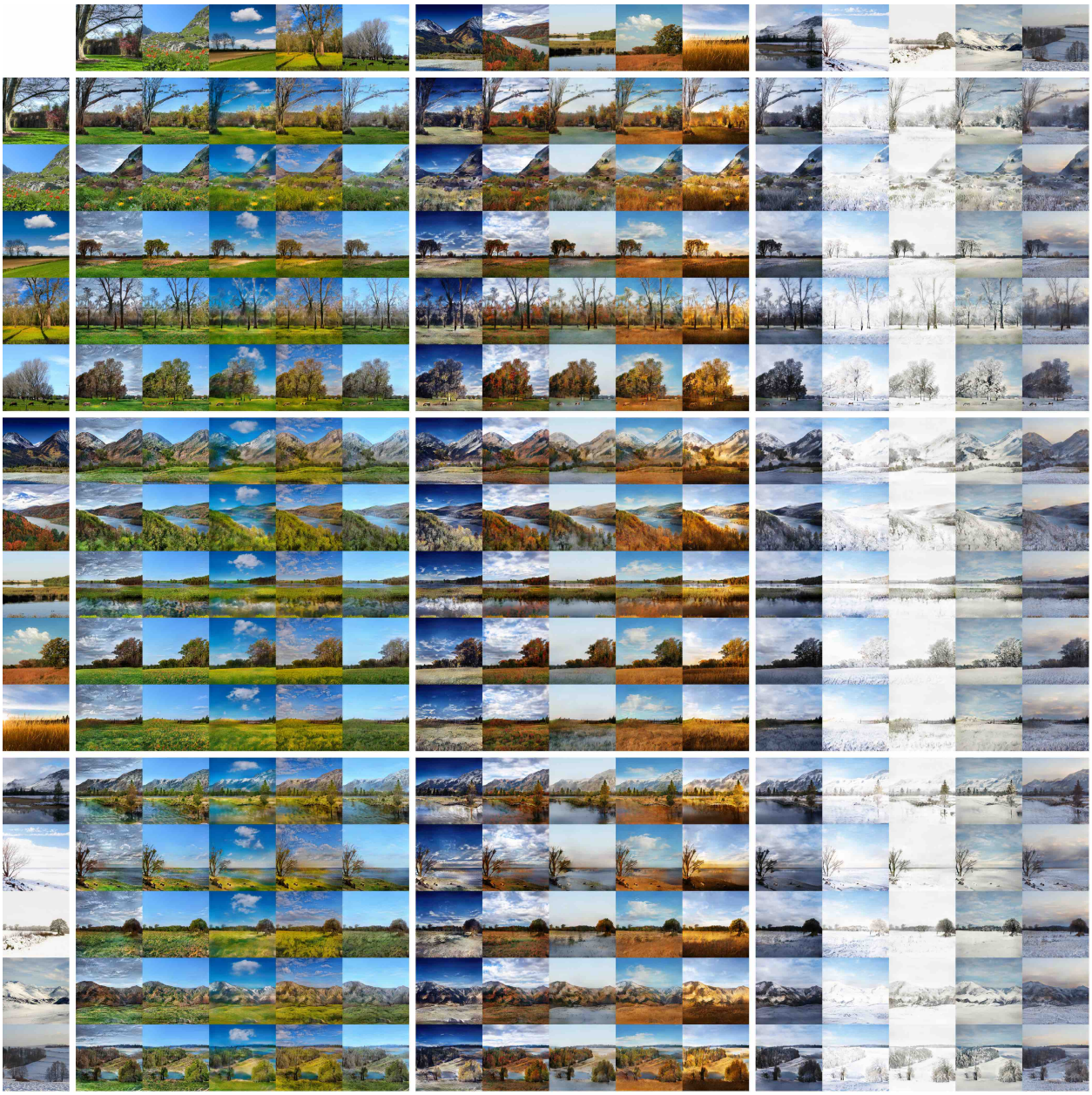

Figure 7: Sensorium is a general purpose image translation network which can synthesize a variety of image tasks despite its simple training objective. Here we illustrate performance on the Seasonal change problem, translating between Spring, Autumn, and Winter.

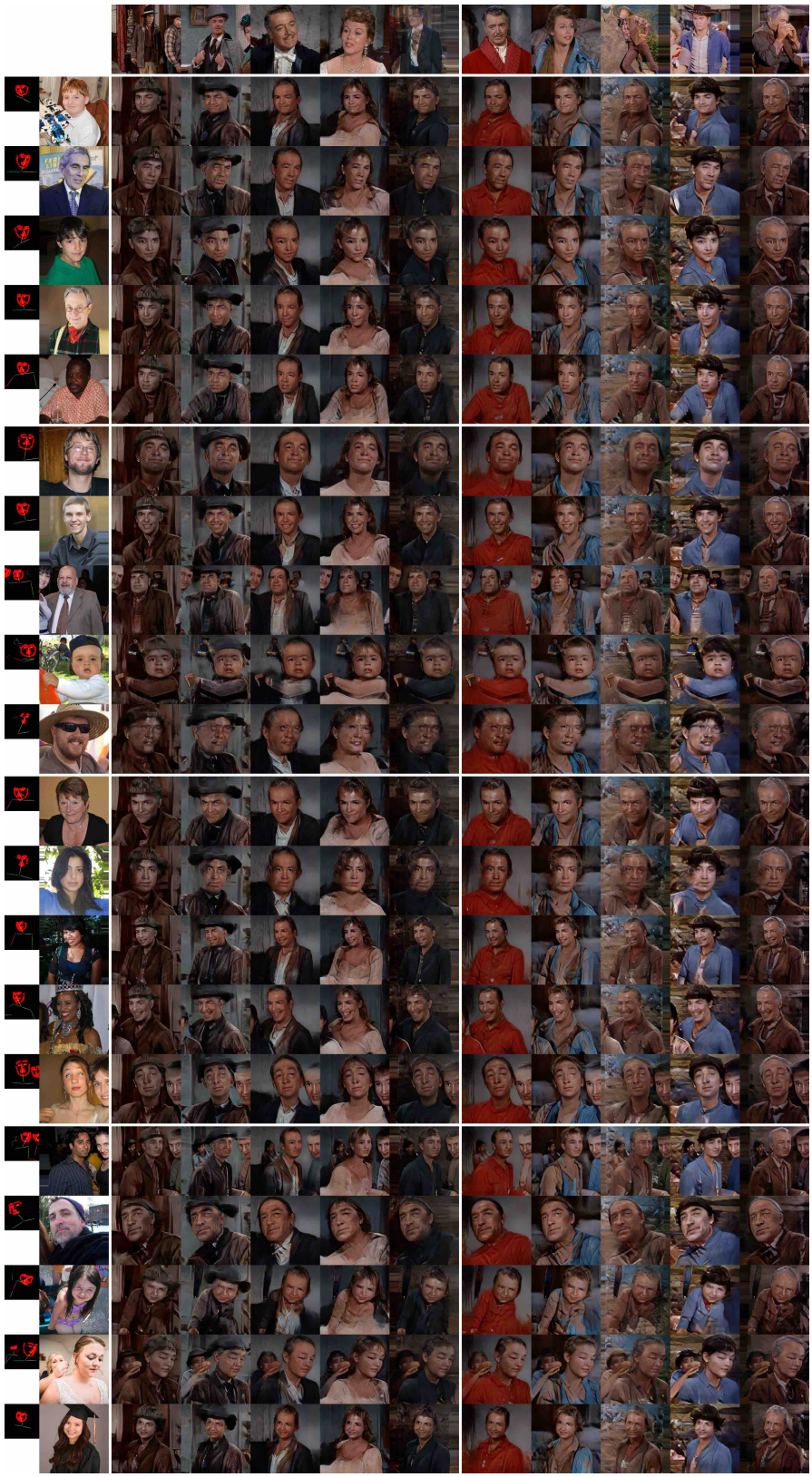

Figure 8: Under our Sensorium approach, the user fixes a content representation via the choice of derived conditioning, and all other features become encoded as style. Here, when translating from FFHQ-Wild to ClassicTV-Bonanza we fix the content representation as KeyPoints. Therefore, the synthesized images pull all else from the style reference images, including hats, clothing, and hair.

References

[1] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 1

[2] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 1